The Development of a Multisource and a Systematized Database for Economic and Policy Impact Analysis

■

Ec. Filippo Oropallo
Cercetător
Institutul Italian Național de Statistică
Ec. Lorenzo Lo Cascio[2]
Consultant pe probleme economice
Grupul META, Italia

Abstract. *ISTAT is involved in various EU projects with the objective of "supporting the Lisbon objectives, EU governance and the process of national policy coverage with the best EU-wide and national policy impact and evaluation analyses".*

Existing knowledge on policy impact analyses is approximate. The "facts" on the impact of policies are charted only at the aggregate level and with a high degree of approximation. Macro indicators have well-known pitfalls and drawbacks. Understanding how policies affect economic performance and developing better indicators to gauge their effects is central to endow the EU with a set of efficient and fair policies. The gap in European knowledge and capacity for Policy Impact Analysis is patent.

The DIECOFIS EU-FP5 project has taken up the challenge of reducing this gap in the field of taxation. Results have been quite encouraging and have open new vistas for future work. Particularly notable has been the development of a system of micro-founded indicators, based on factuals and counterfactuals, estimated through micro-simulation models.

This has led to the current utilization of such a tool in the ex-ante microsimulation of the effects of several reforms of corporate taxation

Key words: *integration problems: enterprise performance; multisource database; microsimulation models; fiscal impact.*

■

Introduction

Micro-founded indicators on enterprise performance and fiscal microsimulation models require a great deal of information. This is normally scattered in different statistical surveys and administrative sources. Each different data source is conceived to serve different purposes and, in many instances, may refer to different units or different definitions may be used for the same unit. Any attempt to bring together data from different sources has to overcome complex problems in terms of sheer access, and integration and systematisation. This paper addresses the following two broad questions:

(i) the reconciliation of administrative and survey sources;

(ii) the integration of sample survey data with exhaustive information from registers and other various administrative sources.

We expose also an analysis of the integration problems that were faced, the architecture of the integration process that has been adopted and how we expect to attain the final integrated information system.

The integration of administrative data (and survey data) can be solved by exact matching techniques, while the integration of data taken from two (or more) surveys can be solved by statistical matching techniques. This is often the case, since surveys rarely contain data on the same enterprises.

The sources

In order to build the integrated and systematized information system on enterprises needed to support economic analysis and for the development of tax microsimulation models and micro founded indicators, the first step is to select the "spine" information that will be use as a basis for the integration process. At ISTAT, the "spine" is constituted by the statistical register of Italian active enterprises (acronym ASIA)[1]. This is the result of an integration process of different administrative sources and represents the best "hanger" for data integration purposes. On this hanger, information from the following sources can be put.

These include:

- Large Enterprise Accounts (Italian Acronym is SCI, cf. ISTAT 2000);
- Small/Medium Enterprise Survey (enterprises with less than 100 workers) (Italian Acronym is PMI, cf. ISTAT 2001b);
- Manufacturing Product Survey (PRODCOM, cf. ISTAT 2001a) and Cost Structure Survey (ISC);
- Foreign Trade Survey (Italian Acronym is COE, cf. ISTAT 2002a);
- Community Innovation Survey (CIS, cf. ISTAT 1999);
- ICT Survey (cf ISTAT 2002b).

All above ISTAT surveys are based on common EUROSTAT standards and classifications (as shown in Figure 1). This implies that the DIECOFIS database can serve to microsimulate the impact of public policies not only in Italy and that a path for the creation of an EU statistical information system has been traced.

The main effort which it was necessary to undertake is about the development of a methodology that allow the data linkage between the information of the above surveys and the whole enterprise universe, represented by the data register on enterprises.

In the ASIA[2] archive, ISTAT files all active enterprises (cf. Eurostat 1999) except for those belonging to Agriculture, Forestry and Fishing (A, B sectors according to NACE classification) and the Public Sector (L, O91, P and Q). This can be used as a starting point or common basis for the linkage of all survey data. In the ASIA archive the following information is included:

- Identifier: internal code, name, fiscal code, vat number, telephone, address;
- Localisation: geographical reference;
- Typology: economic activity and legal form;
- Demographic: status and transformations;
- Size: Turnover and employees.

The information coming from the administrative sources that have been integrated in the DIECOFIS database include:

- The Chamber of Commerce (CA = Commercial Accounts): data from the CC's Annual Report. These complement ISTAT business survey of account system (SCI and PMI) for all corporate, co-operatives and consortium enterprises only.
- The Revenue Agency (FISCAL=Fiscal Data): data from RA annual tax returns.
- The Social Security Administration (SSD = Social Security Data): data from SSA returns.

These two latter sources permit to obtain precise information on tax and social contribution revenues, and thus to calculate the actual tax burden on enterprises, which can be used to test the model's output (e.g. "counterfactuals").

Looking at the quality of the available information, the enterprise size seems to be a "key" variable (Denk, Oropallo, 2003). In fact, exhaustive information (which covers the whole universe) is available for large enterprises that have at least 100 workers, while for small and medium ones we can collect data from a sample of enterprises.

A second characteristic that appears to be very important is the legal form, as the type of tax that an enterprise is required to pay depends on it.

Integration Issues

If we focus the attention on the integration of statistical data with business registers and with administrative data (Giovannini, Sorce, 2001, EUROSTAT, 1999), the first problem we have to

solve is to identify the business unit. This means basically choosing a variable which can be a unique key and act as a natural bridge between the different sources. In almost all firms' databases we have chosen either the VAT code or the fiscal code.

Another important question relates to possible changes to the business (Black, 2001) during the enterprises' life. In fact, the same enterprise may appear as a different unit because of transformation events. Usually two types of changes are considered:

- Changes involving a single unit (changes in kind of business classification, in size or localisation);
- Changes in the number of units (death, birth, divestitures and splits, mergers and acquisitions).

As a consequence of changes or in the presence of new-born firms, we may find that the business register doesn't contain all the units of a survey and it is necessary to distinguish between the case of new firms and that of transformed units. In the latter case, a problem of identifying the successor of the initial business can arise. In some cases, the VAT number of the new unit is different but the fiscal code is the same. A correspondence table containing old and new codes or a table containing the fiscal code and the many VAT numbers used by the enterprise has been used in order to solve this kind of problems.

Integrating different data sources implies the record linkage of two or more files containing units of the same population. This means dealing with both Exact and Statistical matching problems (FCSM, 1980). An example can be found in the use of the PMI survey. Mainly, two problems arise:

- On the one hand, the information contained in the ISTAT PMI survey only refers to profit, loss, staff costs and other variables. No data about enterprise's assets and liabilities is contained.

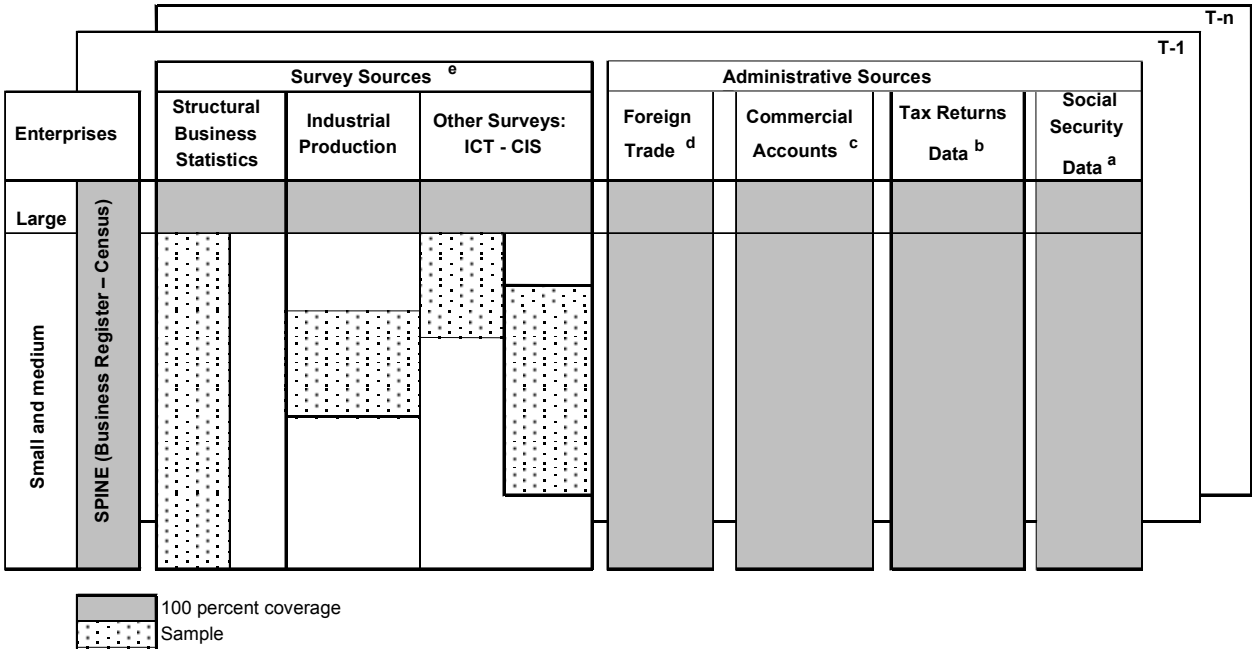
In order to solve this first problem, we need to integrate PMI data with administrative data from the Chambers of Commerce, at least for all enterprises that exist as a separate legal entity, using exact matching techniques, where the key variable is the fiscal code.

- On the other hand, no exact linkage with other Istat surveys is possible, because of Istat's intention to avoid imposing an excessive burden on respondents. ISTAT, in fact, does not include an enterprise in more than one survey in a year.

The resolution of the second problem is possible through statistical matching techniques. By using several different statistical matching techniques, missing variables are drawn from the available information. Missing values are reconstructed with reference to enterprises that have similar characteristics and for which the information is available (FCSM, 1980). The characteristics in question are the size, the economic activity, the geographical area, the legal form etc. The following chart shows the coverage of all sources available. This is the starting point for applying any matching technique.

Nontrivial questions are sampling problems. After merging survey data with exhaustive data taken from a subset of the population, sampling weights have to be recalculated. An example is the merging of variables taken from PMI surveys with variables from foreign trade survey, which are data sets that cover all exporting/importing enterprises. In this case, sampling weights have to reflect the new proportion of export/import enterprises in each stratum of the sample. Sampling re-weighting techniques are also used in the case of changes in a single unit, when it changes its type of business, size or localisation.

Being able to rely on an integrated and systematised database for a sufficient number of years, in this overall systemic perspective, basically means that it would become possible to go beyond overall indicators (measuring "averages") and measures of inequality and dispersions (measuring overall inequality) and to "*slice*", "*dice*", "*drill up"* and "*drill down"*, "*drill through"* and "*drill across*" information *hyper* and *micro cubes*, that is to move horizontally and vertically, across dimensions and over time, and chain link indicators. These *newly-built* indicators would refer to different dimensions and be characterised by systemic features that can be studied to identify factors/areas of weakness or strength; of progress or decline; gains and losses. Accordingly, it would become possible to study aspects relating to structure, composition, distribution and dispersion.



(a) Enterprise with employees only - This database is shaped by the National Social Security
(b) All enterprises - This database is shaped by the National Tax
(c) Incorporated enterprises only (Their account system is regulated by UE
(d) Exporting enterprises only
(e) These sources are governed by: (i) Council Regulations: no 3924/91 - survey of industrial production; (ii) n. statistical units for the analysis of the production system; (iii) no. 58/97 - structural business statistics; (iv) Regulation: no. 1618/99 - evaluation of quality of structural business

Figure 1. *General Framework*

The above figure shows the general framework of the sources to be integrated (Denk, Oropallo, 2002), grey areas represent exhaustive data, dotted areas sample data.

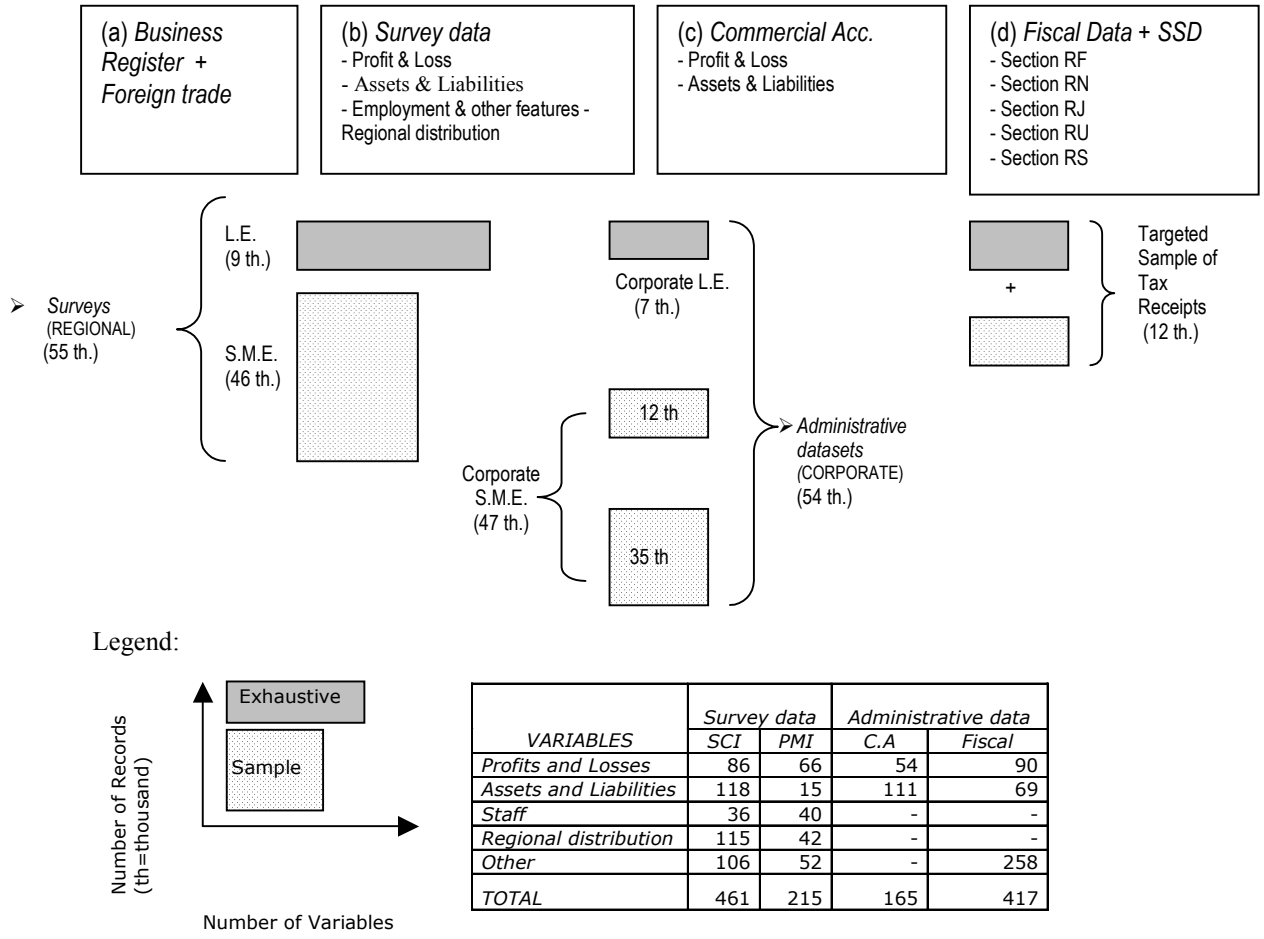
The DIECOFIS Database

First Stage of Integration

The first step of the DIECOFIS database integration was concerned with the creation of the following datasets:

Survey datasets (Regional datasets), which include information on 55 thousand corporate firms, coming from several ISTAT surveys (e.g. Statistical data). Information is exhaustive for large corporate firms (8.8 thousand firms with 100 workers or more). Sample data are available for small corporates (45.9 thousand firms, roughly 1.5 percent of the universe).

Administrative datasets (Corporate datasets), which include information on 54 thousand corporate firms, coming from the Commercial Accounts data (e.g. Administrative data). Information is exhaustive for large corporate firms (7 thousand firms with 100 workers or more). Sample data are available for small corporates (47 thousand firms, roughly 1 percent of the universe).



VARIABLES	Survey data SCI	Survey data PMI	Administrative data C.A	Administrative data Fiscal
Profits and Losses	86	66	54	90
Assets and Liabilities	118	15	111	69
Staff	36	40	-	-
Regional distribution	115	42	-	-
Other	106	52	-	258
TOTAL	461	215	165	417

Figure 2. *First Stage*

Figure 2 shows all separated pieces of data that have been merged with the Business Register and that are described below.

The *Business Register ASIA* (years: 1996 – 2000) includes basic information on the whole universe of Italian active enterprises, so that the contained characteristics can serve as auxiliary variables in the processes of imputation and estimation. These are: (1) Geographical reference, (2) Sector of economic activity, (3) Legal type (4) Dimension (independent workers and employees) and, for a portion of them, the annual turnover. This merging activity makes it possible to proceed with the second stage of integration: the missing data reconstruction, obtained by using matching techniques.

Business Register

Table 1

Year	1996	1997	1998	1999	2000	2001
Large Enterprises	8,091	8,684	8,924	9,240	9,741	10,125
SME	3,862,383	3,761,446	4,040,250	4,122,853	4,212,916	4,287,340
Total	3,870,474	3,770,130	4,049,174	4,132,093	4,222,657	4,297,465

The 1999 *SCI survey* (data are available also for year 1998) contains information about 8.734 enterprises. They refer to the universe of large enterprises with 100 or more workers (with the exclusion of the J division, Financial sector). Among these, we find 7.339 corporate enterprises, about 84 percent of the total.

SCI Survey

Table 2

Year	1998	1999
Corporate Enterprises	7,124	7,339
Non Corporate Enterprises	1,330	1,395
Total	8,454	8,734

The 1999 *PMI survey* (data are available also for year 1998) contains information about 45.947 enterprises, of which 15.329 are corporate enterprises.

PMI Survey

Table 3

Year	1998	1999
Corporate Enterprises	15,372	15,329
Non Corporate Enterprises	32,112	30,618
Total	47,484	45,947

With respect to Commercial Accounts data (years '98-'00), we have, for the year 99, a sample of 53.532 enterprises: 6.902 of these are present in the SCI survey as well; 11.905 enterprises are in the PMI survey too; the remaining 34.725 enterprises are present in the ASIA register but not in any survey.

CA data

Table 4

Year	1998	1999	2000
Large Enterprises	6,145	6,902	5,778
SME Enterprises	48,313	46,630	43,653
Total	54,458	53,532	49,431

The Fiscal Dataset contains a targeted sample of tax returns for the year 1999. It contains all large corporates and a very small sample of small enterprises.

Fiscal Data

Table 5

Year	1999
Large Corporate Enterprises	7,340
SME Corporate Enterprises	4,535
Total	11,875

The Social Security Dataset contains data on all enterprises which have one employee at least.

Social Security Data

Table 6

Year	1999
Large Corporate Enterprises	9,021
SME Corporate Enterprises	1,061,932
Total	1,070,953

In this first stage, other surveys have been linked with the business register:

- The *PRODCOM* survey (ISTAT, 2001a). This survey is exhaustive for large enterprises and there is a sample of small and medium ones (approximately 35,000 units). It covers the manufacturing sector only. Other sectors, such as trade and services, remain uncovered. Moreover, for small and medium enterprises, we have the problem of the missing link between PMI sample units and units from the PRODCOM sample.
- The Foreign trade survey (*COE*), integrates information about foreign trade for the totality of enterprises (ISTAT, 2002). It is derived from custom data and covers all the population of enterprises engaged in foreign trade (approximately 260,000). It contains the value of every item exchanged (with a detail of 8 digits) for each country of destination and origin.
- The Survey of Technological Innovation of enterprises (*CIS* - Community Innovation Survey) collects information on expenses for innovation projects and on the type of innovation in question. The purpose is to estimate the input and output of the innovation process that takes place in enterprises. This survey is led on a representative sample of 5,256 enterprises that are part of the population of industrial enterprises with 20 workers or more. This is not an annual survey but is carried out every 4 years.
- The *ICT* survey (Information and Communication Technologies) tries to gather information on enterprises' use of Information and Communication Technologies and electronic commerce, in order to highlight "new economy" activities. Enterprises with 10 or more workers in the manufacturing sector and in part of the services sector are the reference units. The representative sample contains 7,000 units. The first year of issue is 2001.

Second Stage of Integration

The second stage of integration is synthesized in the chart below.

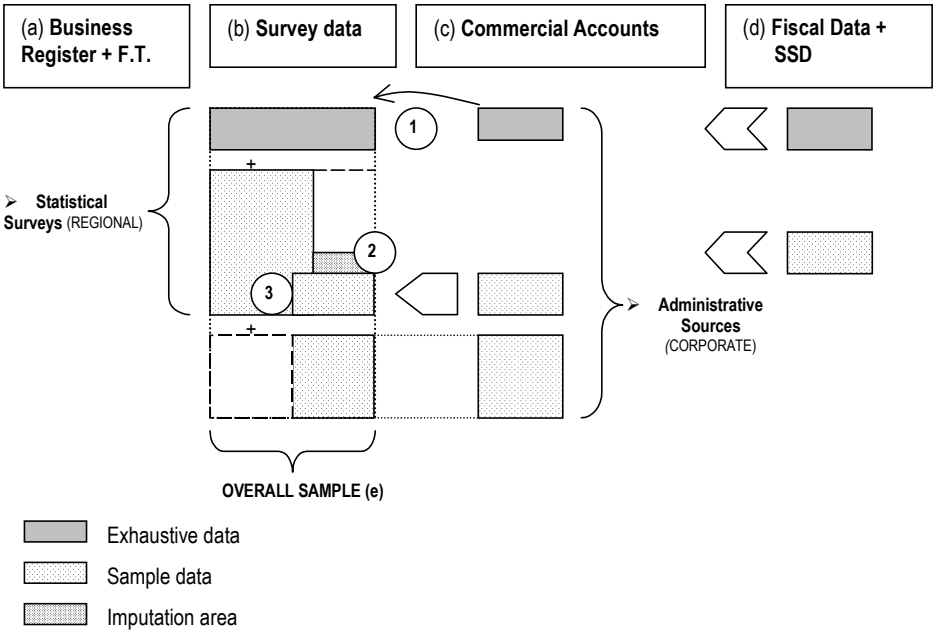


Figure 3. *Second Stage*

In this second step three main issues can be distinguished: (1) the reconciliation of survey data with administrative data, (2) the problem of missing data, (3) the estimation process.

Reconciling Survey and Administrative Data

The harmonization of variable definitions has been, preliminarily, required in order to produce metadata information. When it has been possible, the same variables from different sources have been compared. An example on two important variables is illustrated in the chart below. The comparison of the values of the variables included in both *Corporate* and *Regional* datasets shows a regular distribution of relative differences. Roughly 80 percent of SME under observation have a discrepancy range of ±2 percent, while roughly 50 percent of large enterprises fluctuate between ±2 percent.

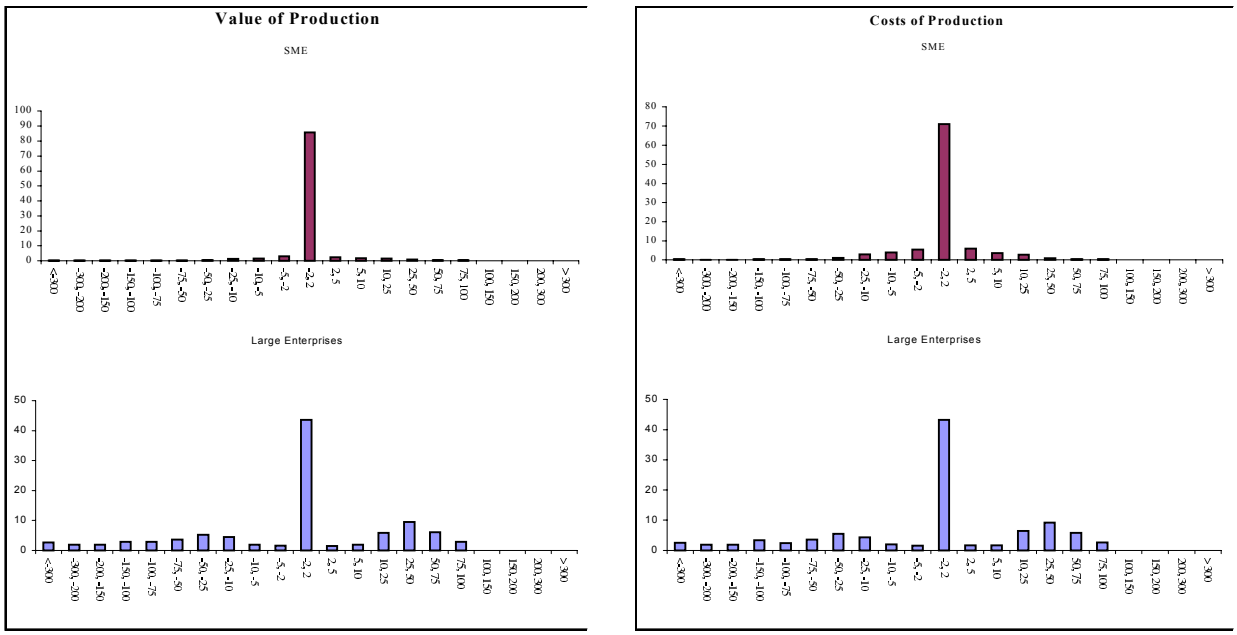


Figure 4. *Comparisons*

Imputation

By integrating the information coming from the Administrative sources, it was possible to reconstruct missing data. In addition, an analysis of discrepancy of variable values has been performed on the units that are present both in the administrative and statistical sources.

CL	Range	SME Enterprises Discrepancies	
		Obs	Perc
1	<-300	10	0.1
2	-300, -200	5	0.0
3	-200, -150	7	0.1
4	-150, -100	22	0.2
5	-100, -75	16	0.1
6	-75, -50	39	0.3
7	-50, -25	157	1.3
8	-25, -10	392	3.3
9	-10, -5	467	3.9
10	-5, -2	664	5.6
11	**-2, 2**	**8449**	**71.0**
12	2, 5	771	6.5
13	5, 10	473	4.0
14	10, 25	323	2.7
15	25, 50	76	0.6
16	50, 75	18	0.2
17	75, 100	8	0.1
18	100, 150	4	0.0
19	150, 200	3	0.0
20	200, 300	1	0.0
21	> 300	1	0.0

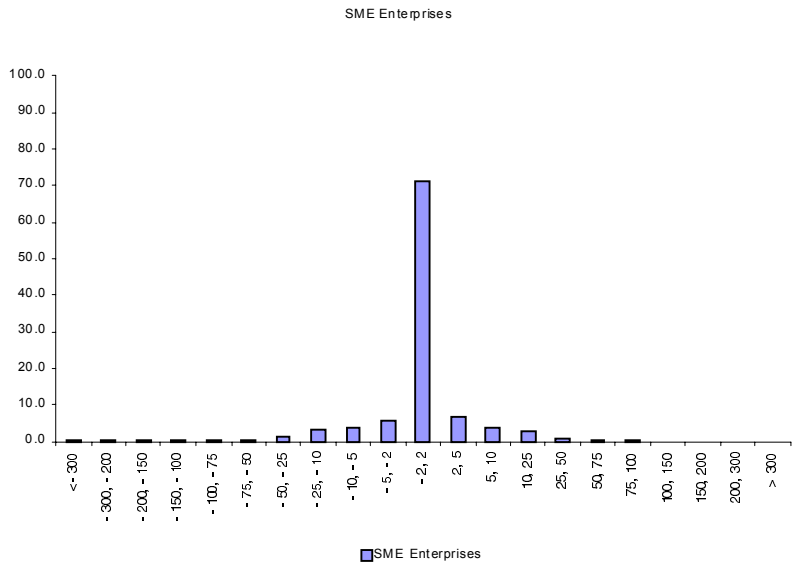


Figure 5. *Analysis of discrepancy*

After the reconstruction of information by integrating survey and administrative sources through a one-to-one match and eliminating outliers, missing data remains where an exact matching has not been possible. In this case, data imputation has been performed through statistical matching techniques. This problem can be viewed as an imputation problem (D'Orazio, Di Zio, Scanu, 2001). In fact, the main difficulty is to choose in two complementary datasets the similar units.

The tables below show on which basis missing data have to be imputed: being 1999 the year of reference and taking the corporate firms of the Regional dataset (15 thousand on a total of 45 thousand firms), 11,905 records are linked with the information contained in the administrative data sources. For the units for which no link has been possible, information has to be reconstructed and missing data has to be imputed (dotted region in figure 3). Several techniques may be applied to achieve this objective.

Merging activities

Table 7

MERGING Regional/Corporate

Cross Sectional integration

YEAR 1999	REGIONAL (99)	of which corporates	Year 1999	
			Linked with Commercial Accounts Dataset	Not linked with Commercial Accounts Dataset
SME Enterprises	45,947	15,329	11,905	3,424
Large Enterprises	8,734	7,339	6,902	437
Total	**54,681**	**22,668**	**18,807**	3,861

Longitudinal integration 1999 – 1998 – 1997 - 1996

YEAR 1999-1998	REGIONAL (99)	of which corporates	Year 1998	
			Linked with Commercial Accounts Dataset (a)	Not linked with Commercial Accounts Dataset (a)
SME Enterprises	45,947	15,329	11,801	3,528
Large Enterprises	8,734	7,339	6,937	402
Totalv	**54,681**	**22,668**	**18,738**	**3,930**

YEAR 99-97-96	REGIONAL (99)	of which corporates	Year 96-97 (dataset regional96_97)	
			Linked with previous surveys datasets (a)	Not linked with previous survey datasets (a)
SME Enterprises	45,947	15,329	4,006	11,323
Large Enterprises	8,734	7,339	6,681	658
Totalv	**54,681**	**22,668**	**10,687**	**11,981**

Imputation methods that have been used in the DIECOFIS database construction include:

- Regression based (EC-JRC, ISTAT, 2003)
- Hot deck (RIDA) (Abbate, 1997), implemented in the software CONCORD (ISTAT – Concord v1.0, 2002c) that operates in SAS environment.
- Multiple imputation (EC-JRC, ISTAT, 2003) also using a SAS procedure known as "PROC MI".

First results of the application of two of the above mentioned procedures are reported below:

Model for Single imputation

Table 8

Variable	Description	Distribution	I/O
X1	Value of Production	Continuous	Input
X2	Total Employment	Continuous	Input
X3	Localisation	Discrete	Input
X4	Business Sector	Discrete	Input
Y	Total Asset (*dependent parameter with missing values*)	Continuous	Output

As regards single imputation, we have identified the regression model Y=f(X1,X2,X4) and Y=f(X1,X2,X3,X4) as the best models (the continuous variables are transformed in logarithmic values).

The *Multiple Imputation* method imputes several values (M) for each missing value (from the predictive distribution of the missing data), to represent the uncertainty about which values to impute. The M versions of completed datasets are analyzed by standard complete data methods and the results are combined using simple rules to yield single combined estimates (e.g., MSE, regression coefficients), standard errors, p-values, that formally incorporate missing data uncertainty. The pooling of the results of the analyses performed on the multiply imputed datasets implies that the resulting point estimates are averaged over the M completed sample points, and the resulting standard errors and p-values are adjusted according to the variance of the corresponding M completed sample point estimates. Thus, the „*between imputation variance"*, provides a measure of the extra inferential uncertainty due to missing data (which is not reflected in single imputation).

The results obtained after 50 imputations for the variable *Y* (total assets) are shown in Table 9. The relative increase in variance (*r*) due to the multiple imputations is very small, indicating that the statistical uncertainty due to missing data, likewise (variation across the imputed datasets) is small.

Estimates of Y (complete data) obtained after m=50 imputations
(Confidence level for interval estimates is 95%)

Table 9

	Estimate	Bl Variance	Wl Variance	Total Variance	Std Error	Low End-Point	High End-Point	r
Y	6.5725	$1.3653\,10^{-6}$	0.7918	0.7918	0.8898	4.8284	8.3165	1.758910^{-6}

Next, we analyse the results considering the regression coefficients as the estimates (Table). In this way we will explore the robustness of the imputed datasets.

Some definitions:

T-ratio is defined as the estimate divided by its standard error (appropriate for testing the null hypothesis that the quantity is equal to zero);

df shows degrees of freedom for Student's t approximation, and

p-value is for testing the null hypothesis that the quantity is equal to zero, against the two-sided alternative hypothesis that it is not zero.

Linear regression coefficients (dataset after imputation) obtained after m=50 imputations (Confidence level for interval estimates is 95%)

Table 10

	Estimate	Standard error	t-ratio	Df	p-value	Low End-Pnt	High End-Pnt	r
B0	1.83695	0.0319513	57.49	767	0.0000	1.77423	1.89967	0.3380
B1	0.655848	0.00611069	107.33	795	0.0000	0.643853	0.667843	0.3302
B2	0.269159	0.00706406	38.10	1154	0.0000	0.255299	0.283019	0.2595
B3	-0.0121438	0.00675127	-1.80	1949	0.0722	-0.0253843	0.00109661	0.1884
B4	0.0654595	0.00542086	12.08	3033	0.0000	0.0548306	0.0760884	0.1456

A comparison of the *t-ratio* and the *t-distribution* for *B3* reveals that the null hypothesis (*B3*=0) is true.

This first test on the reconstruction of missing information is made through sensitivity analysis (EC-JRC, ISTAT, 2003). With single and multiple imputation techniques alternative estimates are obtained and a test is produced to validate the better model in the single imputation (table 10) and to accept multiple imputation results.

Weight Adjustment

Weight adjustment concerns the re-calculation of the weights of the sample units in order to produce reliable estimations, even after the integration process. The variables usually used at ISTAT for sample stratification are NACE (four digits), 21 Regions and 5 employment classes. In this case, there is the need to add a new dummy variable, corporate/non corporate firm. The technique that is used is the Generalized Estimation Method (Falorsi, 1995), where calibrated estimators are applied. Weights are adjusted through the minimization of a distance function between the initial and final weights. The distance function is subject to two boundary conditions: 1) the sum of weighted firms in each stratum has to coincide with the sum of register firms; 2) the sum of weighted employment in each stratum has to coincide with the sum of register employment (Estevao, Hidiroglou, Sarndal, 1995, Falorsi, 1995, Ballin, Falorsi, Pallara, 2000).

First results are produced by considering the dummy variable "corp" (0 = non corporate; 1 = corporate) and one only constraint, i.e. the total number of enterprises. In this case, the solution is the following:

$$Weight\_new_{ik} = weight_{ik} * Total\_Population_{k} / Sum\_of\_weights_{k.}$$

Where: i=enterprise and k=stratum.

At this stage, a first re-weighting procedure has been run, in order to correctly re-calculate sample weights, so that the sum of corporate weights is equal to the number of corporate enterprises. When there is a higher number of constraints, the problem of constrained optimization has to be handled with the GENESEES software (Istat, 2002d).

Overall Database for Policy Impact Analyses

The final result of the integration process is the overall dataset, which is representative of the universe of enterprises. Data marts are extracted from this database to serve fiscal microsimulation analysis and to produce systemic analyses.

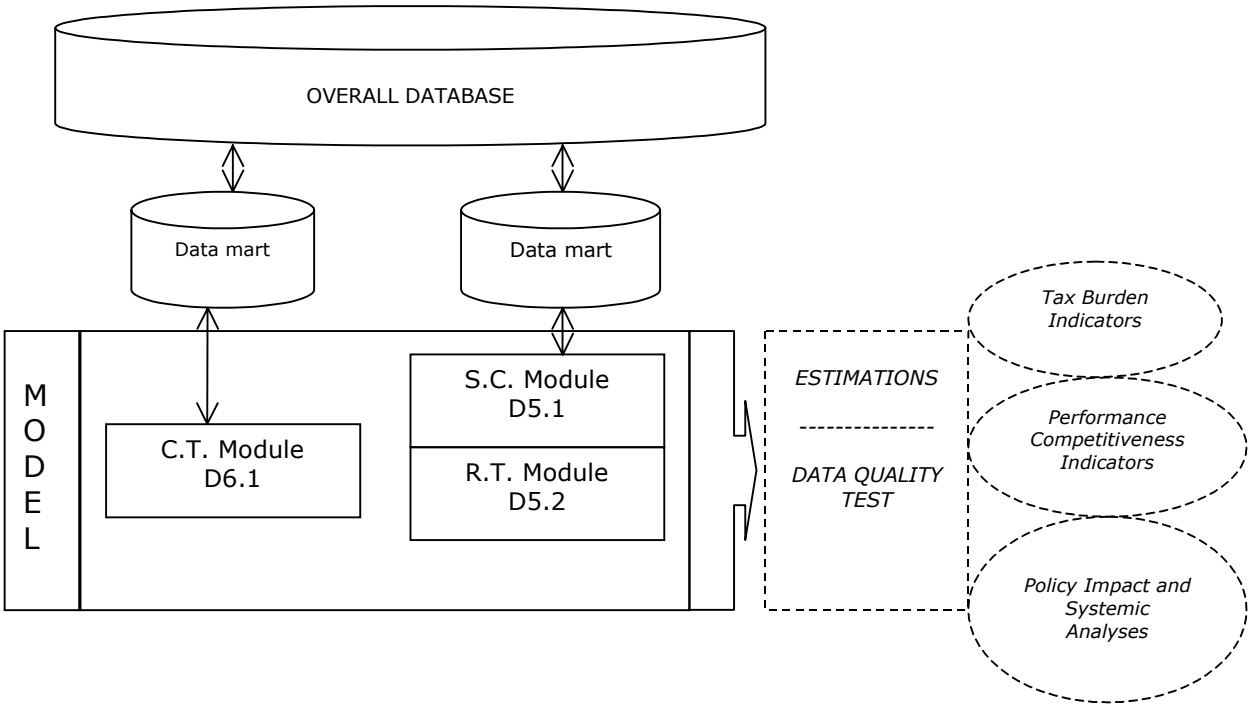


Figure 6. *Overall and P.I.A.*

From the integrated data base the model should estimate the taxable yield for every type of tax. Even if ISTAT surveys do not cover all data needs for a precise calculation of taxable incomes, of the voices of deductions of taxable income and of deduction of the tax, we can, however, make hypothesis on the behaviour of the enterprise or select proxy variables.

The DIECOFIS Microsimulation Model

Social Security Contributions Module[3] and Regional Tax Module[4]

The Social Contributions module that has been developed needs data on the cost of staff for every type of worker. The taxable amount is the yield of the enterprise (for single and the associates) or the gross wage. The contributions share is differentiated according to professional category (manager, employees, etc.) and type of contract (formation, collaborations, etc.).

The equation will be:

$$CS = \sum_i tc_i \times w_i$$

Where t_{ci} is the contribution share for the category "*i*" and w_i the wage. In the case that these details for precise calculations of the social security variables are not available, an estimation of the average share can be obtained from the total cost of the social contributions of the enterprise.

The calculation of the Regional Tax (IRAP) on Added Value is carried out in the following way:

IRAP = tr (VP - CP - TO)

Where:

tr represents the share proportion of the regional tax.

VP = Value of Production: Income from sales, variations of stocks, other income.

CP = Costs of Production: Raw materials and consumables, other external charges, value adjustments, amortisation.

TO = Other Deductions: INAIL (National Institute of Insurance Against Accidents at Work) Contributions, apprentices' costs, formation contract jobs and costs for disabled persons.

The illustrated method of calculation "from the top" could be replaced by a calculation method beginning "from the bottom", that is from the Profit or loss and adding the costs of the staff, the devaluation of credits, provisions put aside, and extraordinary and financial income or charges.

Corporate Tax Module[5]

The Corporate Tax module uses balance sheet data and profit and loss data in order to calculate the tax burden on companies. The equation of Italian corporate tax is the following:

$$IRPEG = t_g(U + T_{IND.} + Cr + Crd - PP)$$

Where:

t_g is the legal rate, U is the profit, T_{IND} is the amount of non deductible taxes, Cr tax credits in profit and loss scheme, Crd Tax credit to share dividend and PP the amount of loss brought forward.

Taxes on Corporations however are accompanied by measures that try to reduce the burden. This is the case of the Dual Income Tax and of the measures introduced by the Tremonti reform. In synthesis the variables of interest for their calculation are:

Variables for Corporate Tax Calculation

Table 11

Deduction Tremonti	Dual Income Tax	Special Visco
Int = Investments minus handovers *Mi* = Average (5 year) Investments	ΔE = Variation of Equity	*Inv* = Investments minus amortization

This module adds the fiscal estimation variable that affects corporations and others.

The chart below shows the structure of the microsimulation model.

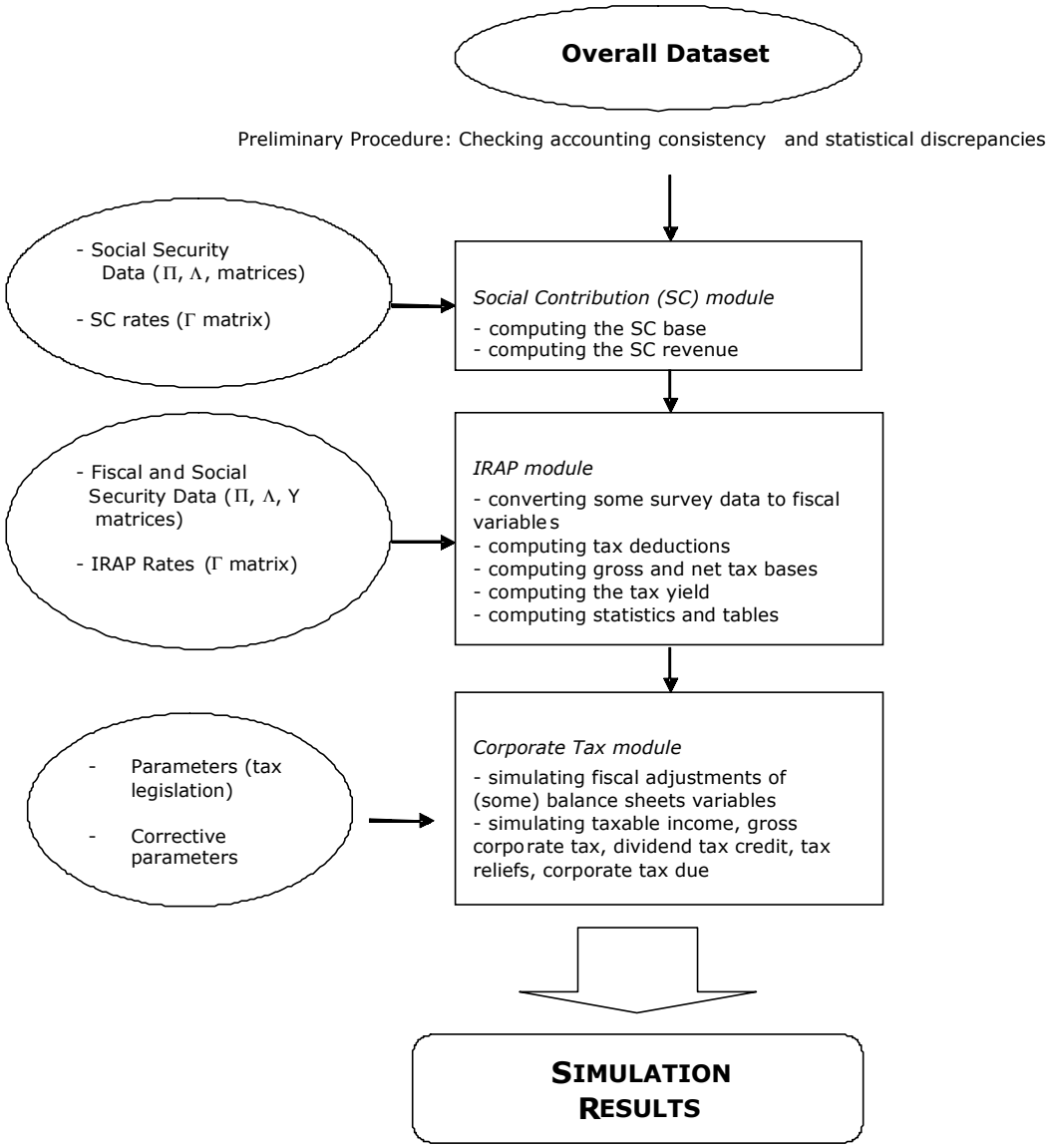
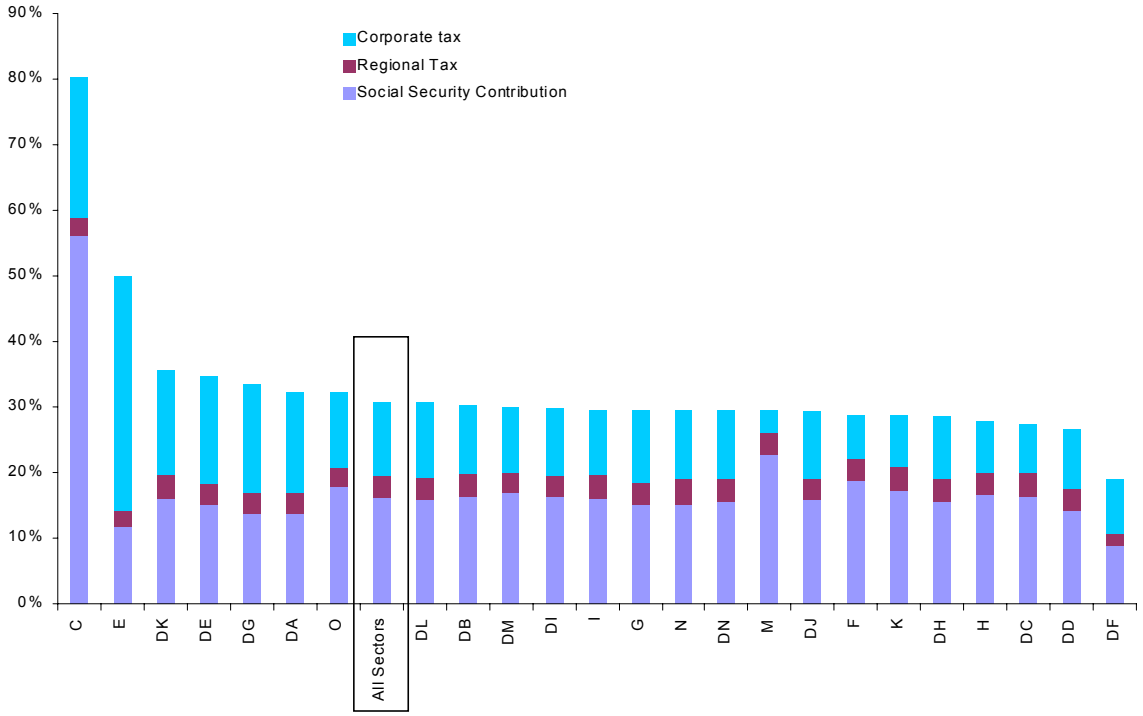


Figure 7. *Microsimulation Model*

Tax Indicators[6]

In order to quantify the direct tax burden on firms, Effective Tax Rates (ETR) and ex post implicit tax rates (EPITRs) have been calculated. First results are produced only for large corporations (more than 8.000 firms). Figure 8 shows the total impact of taxes on the different sectors and illustrates how different taxes concur in forming total fiscal burden.



Source: Diecofis Project's Estimations.

Figure 8. *Effective Tax Burden by NACE sector*[7]

EPITRs are computed using turnover as a denominator. It is used, in the microsimulation exercise, to compare the baseline 1988 with two scenarios: (1) intermediate reform; (2) full reform.

As regards the effects of the full reform, the estimations show some interesting findings. First, the implicit rate drops in all sectors because of the full reform, except "Electricity, gas, water supply" which records an increase of the implicit rate of 0.07 percentage points mainly due to the reform of corporation tax. This is somewhat an expected finding as, according to our simulations, this is the sector where (large) firms seem to have been most favoured by the DIT system. We also recall that the same result applies to companies of the sector "Transport and communication", although for this sector changes to IRAP provide for a reduction of the overall implicit rate. On the whole, falls of the implicit rates are greater for companies in the commerce and in the services sector; the highest reduction is recorded for the sector "Education" (-1,65).

EPITRs for different scenarios by NACE

Table 12

	Scenarios			Differences			
Sector of Activity (Nace)	Baseline 1998	2001	Full Reform 2003	2001-1998	2001-1998	2003-2001	2003-2001
				Overall	Corporation tax	Overall	Corporation tax
Manufacturing, mining (C-D)	4.13	3.84	3.62	-0.30	-0.30	-0.21	0.00
Electrical energy, gas, steam, water (E)	2.79	2.36	2.43	-0.43	-0.43	0.07	0.31
Construction (F)	3.39	3.19	3.00	-0.20	-0.20	-0.18	0.11
Wholesale and retail trade services (G)	2.37	2.22	2.13	-0.15	-0.15	-0.09	0.03
Hotel and restaurant services (H)	4.77	4.60	4.07	-0.17	-0.17	-0.53	-0.08
Transport, storage, comm. (I)	6.79	6.24	5.76	-0.54	-0.56	-0.49	0.29
Real estate, renting and business (K)	6.00	5.88	4.89	-0.13	-0.13	-0.99	-0.21
Education services (M)	4.50	4.47	2.82	-0.04	-0.05	**-1.65**	-0.13
Health and social services (N)	6.32	6.09	5.35	-0.24	-0.24	-0.74	0.08
Other social services (O)	6.05	5.71	5.26	-0.34	-0.35	-0.44	-0.05
Total	4.52	4.24	3.89	-0.28	-0.28	-0.35	0.01

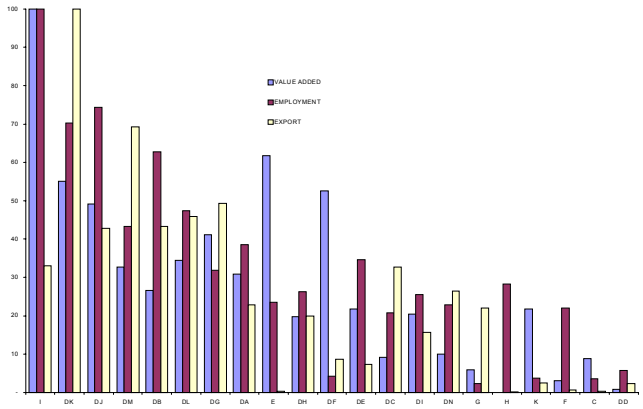
Source: Diecofis Project's estimations.

Composite and Decomposable Indicators

Micro founded indicators ought to be based on a comprehensive and multi source database that embraces all aspects of enterprise features. They must be opportune or scope fulfilling and must fit policy objectives (purpose oriented). They must possess the "well-behaved" and "consistent" properties to be able to support the decision-making process. Among very important indicator properties, the "decomposablility" and the "multi-dimensional" properties are key to better understand social and economic phenomena. Using Istat's integrated data base, a set of performance indicators possessing the above mentioned properties was developed through the GINI index features, each covering a single dimension of firm performance; in a second step, these were aggregated into a composite indicator. The latter permitted a comparison of performance levels of business sectors, geographical areas etc.

The chart below shows some results. The decomposition of the GINI index permits to separate "between classes" effects. Moreover, the weighted sum of such effects produces a composite indicator which leads to a ranking of business sectors.

Three dimensions of enterprises' performance:
(1)Value Added (2)Employment (3)Exports



Composite Indicator

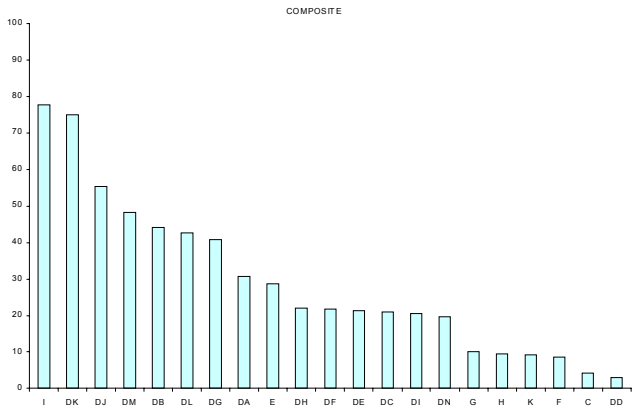
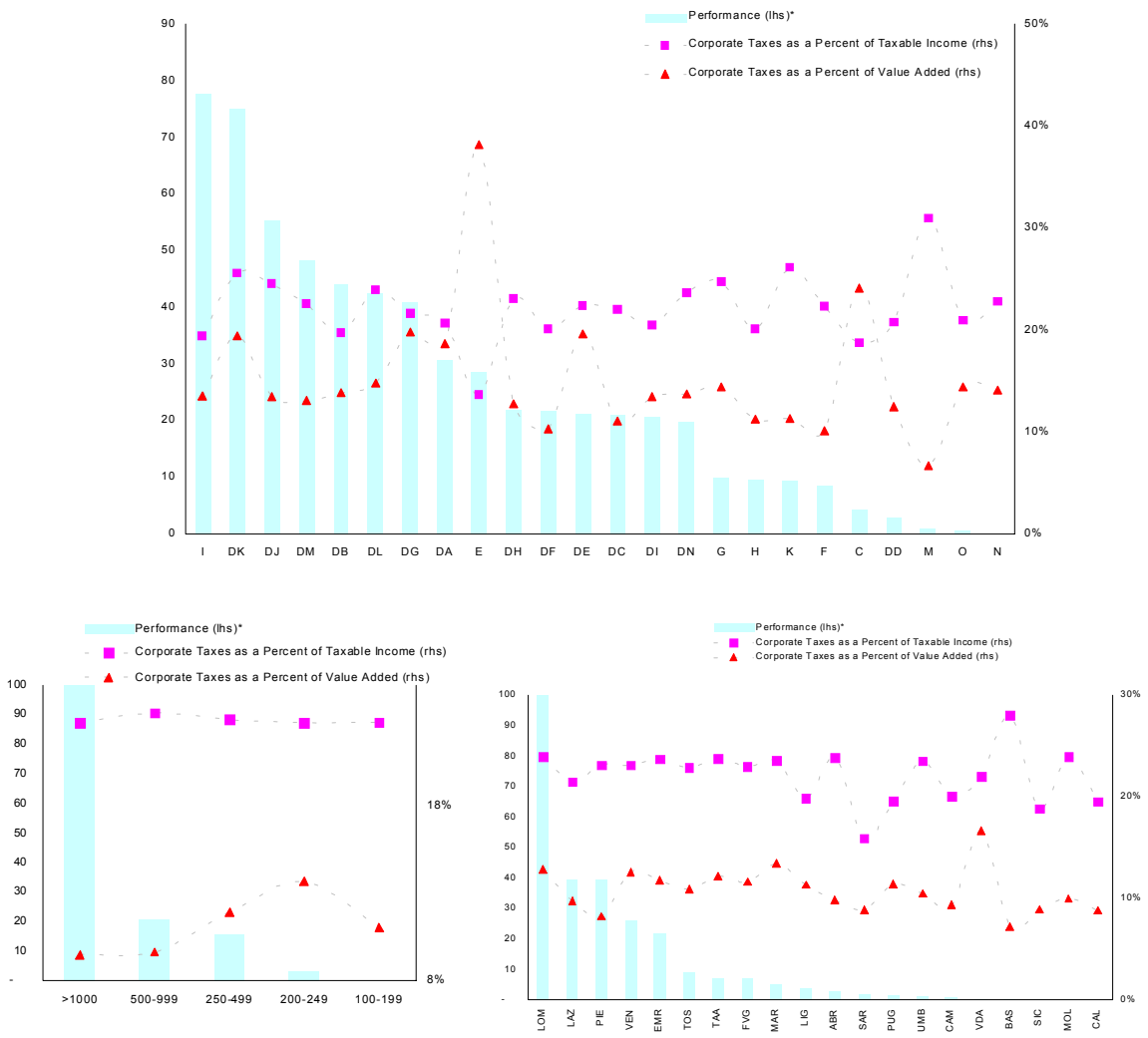


Figure 9. *Decomposition of the GINI between component by economic activity class*



Source: Diecofis Project's estimations.

Figure 10. *Composite Indicator and Corporate Tax Burden, by NACE, size and Region*

Conclusions

In the context of the DIECOFIS project aimed at the development of a system of indicators on competitiveness and fiscal impact on enterprises performance, one preliminary issue is the integration of all available information on enterprises into one multi-source database, so that as many (sensible) data as possible may contribute to the indicators measuring fiscal impact.

The state of enterprise data integration at ISTAT has been briefly summarized. Basically, two integration problems have been identified: the integration of administrative data (and survey data), which may be solved by exact matching techniques, and the integration of data of two (or more) surveys, which may be solved by statistical matching techniques, as, usually, surveys do not contain data on the same enterprises to reduce respondent burden.

The main integration problem at ISTAT with respect to business data is to merge two sample surveys when firms are not the same. In such cases, statistical matching techniques should be applied so that similar units in two samples can be identified, where one dataset is assumed to be the base set and to each base record the value of one or more variables contained in the other dataset gets assigned.

To apply these statistical matching techniques, matching variables are required: one quite apparent option is to use firm characteristics as provided by the business register. These include: (i) Identifiers: ASIA ID, name, VAT number, fiscal code; (ii) Localization: geographical references like region, province, municipality; (iii) Typology: economic activity (NACE) and legal form; (iv) Demographic information: date of birth; (v) Size: number of employees and turnover.

Identifiers will only be used to merge survey data and register data to obtain the other firm characteristics for the enterprises in the survey. Then, those variables may be used for statistical matching. Possibly, location and typology characteristics together with age and size intervals should be used to set up equivalence classes, and then age and size of enterprises should be considered in the calculation of the distance / similarity of base records and records in the appropriate equivalence class of the reference dataset to determine those records to be matched.

The complete development of a user friendly interface and the management of confidentiality issues through the provision of remote access to microdata will allow the user to test, run and validate microsimulation models and to produce new indicators directly on the DIECOFIS integrated and systematised database.

Note

(1) Basically, the actual enterprise state of activity is presumed by means of a logistic model, where the probability of existence is a function a various signs of life, drawn from different administrative sources.

(2) The ASIA project started in 1995, its goal is to improve and update the register of all Italian enterprises. It is the result of the integration of external sources with ISTAT Archives (old Sirio-nai archive, 7 Industry Census and survey SK). External sources are:
- VAT Register of the Ministry of Finances;
- Chambers of Commerce;
- INAIL (National Institute of Insurance Against Accidents at Work);
- INPS (National Social Security Institute);
- Yellow Pages.

(3) See references Bardazzi et al., 2003

(4) See references Bardazzi et al., 2002

(5) See Castellucci, et al., 2002

(6) See Bardazzi et al., (2003)

(7) See the next table

NACE	DESCRIPTION	Firms %	Empl. %	NACE	DESCRIPTION	Firms %	Empl. %
C	PRODUCTS FROM MINING AND QUARRYING	0.2	4.6	DK	MACHINERY AND EQUIPMENT N.E.C.	10.7	2.7
DA	FOOD PRODUCTS, BEVERAGES AND TOBACCO	4.5	2.9	DL	ELECTRICAL AND OPTICAL EQUIPMENT	6.6	3.6
DB	TEXTILES AND CLOTHING INDUSTRY PRODUCTS	8.6	2.1	DM	TRANSPORT EQUIPMENT	4.0	6.9
DC	LEATHER AND LEATHER PRODUCTS	2.2	1.7	DN	OTHER MANUFACTURED GOODS N.E.C.	2.9	1.8
DD	WOOD AND PRODUCTS OF WOOD AND CORK (EXCEPT FURNITURE)	0.8	1.4	E	ELECTRICAL ENERGY, GAS, STEAM AND WATER	0.7	20.4
DE	PULP, PAPER AND PAPER PRODUCTS; RECORDED MEDIA; PRINTING SERVICES	3.5	2.4	F	CONSTRUCTION WORK	3.9	2.1
DF	COKE, REFINED PETROLEUM PRODUCTS AND NUCLEAR FUEL	0.4	6.3	G	WHOLESALE AND RETAIL TRADE SERVICES	8.8	3.3
DG	CHEMICALS, CHEMICAL PRODUCTS AND MAN-MADE FIBRES	4.9	3.6	H	HOTEL AND RESTAURANT SERVICES	2.4	3.6
DH	RUBBER AND PLASTIC PRODUCTS	3.5	2.3	I	TRANSPORT, STORAGE AND COMMUNICATION SERVICES	4.7	13.2
DI	OTHER NON METALLIC MINERAL PRODUCTS	3.6	2.5	K	REAL ESTATE, RENTING AND BUSINESS SERVICES	9.4	2.4
DJ	BASIC METALS AND FABRICATED METAL PRODUCTS	9.1	2.3	M - N – O	EDUCATION – HEALTH AND SOCIAL - OTHER SOCIAL SERVICES	4.6	6.0

References

Abbate, C. (1997), „Completeness of Information and Imputation from Donor with Minimum Mixed Distance “, Quaderni di Ricerca ISTAT, n. 4/1997, pp. 68-102

Alvey, W., Jamerson, B. (eds.) (1997). *Record Linkage Techniques*, Washington, DC: Federal Committee on Statistical Methodology (FCSM)

Alworth, J. S., Castellucci, L. (1993), Chap. 6, Italy, in Jergenson D. W.- Landon, R. (eds), *Tax Reform and the Cost of Capital. An International Comparison, The Brookings Institution*, Washington D.C.

Ballin, M., Falorsi, P.D., Falorsi, S., Pallara, S. (2000). „Il trattamento delle mancate risposte totali nelle indagini ISTAT sulle Famiglie e sulle Imprese (Analysis of total non-response in ISTAT Surveys of Families and Firms)”, *ISTAT Methodological Studies,* 2000

Bardazzi, R., F., Pazienza, M.G., Parisi, V. (2003), *The Effects of the Italian Tax Reform on Corporations: a Microsimulation Approach*. Available on the website http://www.istat.it/diecofis.

Bardazzi, R., Gastaldi, F., Pazienza, M.G. (2002). *The IRAP module, Deliverable 5.2 of Diecofis Project*. University of Florence. Available on the website http:// www.istat.it/diecofis.

Bardazzi, R., Gastaldi, F., Pazienza, M.G. (2003), *The Social Contribution Module, Deliverable 5.1 of Diecofis Project*. University of Florence. Available on the website http://www.istat.it/diecofis.

Belin, R., Rubin, D.B. „A Method for Calibrating False-Match Rates in Record Linkage”. *JASA* 90 vol. 430, 1995, pp. 694–707

Black, J., „Changes in Sampling Units in Surveys of Businesses”. *FCSM Research Conference Papers*, US Census Bureau, 2001

Bontempi, M.E., Giannini, S., Guerra, M.C., Tiraferri, A. (2001). *Incentivi agli investimenti e tassazione del reddito di impresa: una valutazione delle recenti innovazioni normative, mimeo*, available on the website http://www.capp.unimo.it.

Bordignon, M., Giannini, S., Panteghini, P. „Reforming Business Taxation: Lessons from Italy?”, in *International Tax and Public Finance*, vol. 8, n. 2, 2001

Bosi P., Guerra M.C. (2002), *I tributi nell’economia italiana*, Il Mulino, Bologna

Brick J.M., Kalton G., „Handling Missing Data in Survey Research”, *Statistical Methods in Medical Research*, vol. 5, 1996, pp. 215-238

Calza M.G., Inglese F. „Assessment of different approaches for the integration of sample surveys”, ISTAT, 2003, Available on the website http://www.istat.it/diecofis

Castellucci, L., Coromaldi, M., Parisi, V., Perlini, L., Zoli, M., „Report describing country IT Corporate Tax Model and methodology”, *Deliverable 6.1 of Diecofis Project*, 2002, University of Rome Tor Vergata. Available on the website http://www.istat.it/diecofis.

Dempster, A.P., Laird, N.M., Rubin, D.B. „Maximum Likelihood from Incomplete Data via the EM Algorithm”, *JRSS B* 39, 1977, pp. 1–38

Denk, M., Oropallo, F. (2002). *Overview of the Issues in Longitudinal and Cross-Sectional Multi-Source Databases*, www.istat.it/diecofis.

Denk, M. (2002). *Statistical Data Combination: A Metadata Framework for Record Linkage Procedures*. Dissertation Thesis, Dept. of Statistics, University of Vienna

Denk, M., Froeschl, K.A., „The IDARESA Data Mediation Architecture for Statistical Aggregates”, *Research in Official Statistics* vol. 3 (1), 2000, pp. 7–38

Deville J. C., Särndal, C. E., “Calibration Estimators in Survey Sampling”, *Journal of the American Statistical Association*, vol. 87, 1992, pp. 376-382

EC-JRC, ISTAT (2003). *Software analysis - Development of methodologies and of a software for the measurement of statistical quality, and for comparing the robustness of alternative multi-source, integrated databases* – Deliverale 3.1 - DIECOFIS http:// www.istat.it/diecofis/deliverable_list.htm.

Estevao, V., Hidiroglou, M. A., Särndal, C. E., „Methodological Principles for a Generalized Estimation System at Statistics Canada”, *Journal of Official Statistics*, vol. 11, n. 2, 1995, pp. 181-204

D’Orazio, M., Di Zio, M., Scanu, M., „Statistical Matching: a tool for integration data in National Statistical Institutes”, paper ISTAT for NTTS 2001 – ETK 2001 – Crete Conference 18-22 June 2001, Eurostat, JRC(ISIS)

Estevao, V., Hidiroglou, M.A., Sarnald, C.E., „Methodological Principles for a Generalized Estimation System at Statistics Canada”, *Journal of Official Statistics vol.* 11 (2), 1995, pp. 181-204

EUROSTAT (1999). *Use of Administrative Sources for Business Statistic Purposes: Handbook on Good Practices* – Theme 4 (Industry, Trade and Services), EUROSTAT Edition.

Falorsi, P.D., Falorsi, S., „Un metodo di stima generalizzato per le indagini sulle famiglie e sulle imprese (A Generalized Estimation Method for Surveys of Families and Firms)", *Quaderni CON PRI*, 1995, University of Bologna.

FCSM – Federal Committee on Statistical Methodology, „Report on Exact and Statistical Matching Techniques", *Statistical Policy Working Paper* 5, Washington, DC: U.S. Department of Commerce

Fellegi, I.P., Sunter, A.B., „A Theory for Record Linkage", *JASA* nr. 64, 1969, pp. 1183–1210

Froeschl, K.A. (1997). *Metadata Management in Statistical Information Processing*. Wien–Berlin: Springer.

Froeschl, K.A. (1999a), „On Standards of Formal Communication in Statistics", *Working Paper* No. 16, UN-ECE/METIS, *Work Session on Statistical Metadata*, 12 pp

Froeschl, K.A. (1999b), „Metadata Management in Official Statistics An IT-based Methodology Approach", *Austrian Journal of Statistics,* vol. 28 (2), 1999b, pp. 49–79

Froeschl, K.A., Grossmann, W., „The Role of Metadata in Using Administrative Sources", *Research in Official Statistics* vol. 3 (1), 2000, pp. 65–82

Garcia-Molina, H. et al., „Intelligent Integration of Information (I3)", *DARPA-Progress Report*, 1995

García-Solaco, M., Saltor, F., and Castellanos, M. (1996). *Semantic Heterogeneity in Multidatabase Systems*, In: Bukhres, O.A. and Elmagarmid, A.K. (eds) *Object-Oriented Multidatabase Systems*. Englewood Cliffs, NJ: Prentice Hall, pp. 129–202

Giovannini, E., Sorce, A., „Integration of Statistical (survey) data with registers (administrative) data", Paper contributed to the *Meeting on the Management of Statistical Information Technology, 2001*

Informer SA (2003a) - Code for user interface – Deliverable 2.4 – *Diecofis project* – http://www.istat.it/diecofis/deliverable_list.htm.

Informer SA (2003b) – Software User Manual – Deliverable 2.5 – *Diecofis project* – http://www.istat.it/diecofis/deliverable_list.htm.

Inmon, W. H. – *Data Marts and Data Warehouse: Information Architecture for the Millenium* – Informix Corporation.

ISTAT, „L'innovazione tecnologica nelle imprese (Firms' Technological Innovation)", *Note Rapide – July 1999* (LE Survey) - http://www.istat.it/Imprese/Ricerca-e-/index.htm.

ISTAT, „I risultati economici delle medio-grandi imprese Anni 1998-99" (Economic Outcomes of Medium-Large Size Enterprises) - *Statistiche in breve - July 2000* (LE Survey) - http://www.istat.it/Imprese/Struttura-/index.htm.

ISTAT (2001a). *Indagine Prodcom (prodcom Survey)* – Indagine sulla struttura dei Costi (Cost structure Survey) http://www.istat.it/Imprese-e-/index.htm.

ISTAT (2001b) *Struttura e competitività del sistema delle imprese industriali e dei servizi nel 1998 (Structure and competitiveness of industrial and service enterprise system in 1998). - Statistiche in breve - Luglio 2001* (LE & PMI Survey) http://www.istat.it/Imprese/Struttura-/index.htm.

ISTAT (2002a) *Indagine sul Commercio Estero* (Foreign Trade Survey). Current version available at http://www.coeweb.istat.it/.

ISTAT (2002b). *L'uso delle tecnologie dell'informazione e della comunicazione nelle imprese* (The use of ICT in Italian firms) – *Statistiche in breve* http://www.istat.it/Imprese/Ricerca-e-/index.htm

ISTAT (2002c). CONCORD v1.0 - (Generalized Data Editing Software) SOFTWARE GENERALIZZATO PER IL CONTROLLO E LA CORREZIONE DEI DATI RILEVATI NELLE INDAGINI STATISTICHE – MPS – ISTAT 2002 - http://www.istat.it/Metodologi/index.htm.

ISTAT (2002d). GENESEES v1.0 - (GENEralised software for Sampling Estimates and Errors in Surveys) SOFTWARE PER IL CALCOLO DELLE STIME E DEGLI ERRORI CAMPIONARI– MPS – ISTAT 2002 - http://www.istat.it/Metodologi/index.htm.

Jaro, M.A. (1978). *UNIMATCH: A Record Linkage System, User's Manual*, Washington DC: U.S. Bureau of the Census

Jaro, M.A., Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *JASA* 84, 1989, pp. 414–420

Kadane, J.B., „Some Statistical Problems in Merging Data Files", In: *1978 Compendium of Tax Research*, US Dept. of the Treasury, 1978, pp. 159–171 (Reprinted in *Journal of Official Statistics,* vol. 17 (3), 1978, pp. 423–433)

Kalton G., Kasprzik D., „The treatment of missing survey data", *Survey Methodology*, vol. 12, n. 1, 1986, pp. 1-16

Kalton G., Kasprzik D., „Imputing for missing survey responses", *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1982

Kamakura W. A., Wedel M., „Statistical Data Fusion for Cross-Tabulation", *Journal of Marketing Research*, vol. 34, 1997, pp. 485-498

Kovar, J.G., Whitridge, P.J. (1995). *Imputation of Business Survey Data*. In: Cox et al. (eds.), *Business Survey Methods*, New York: J. Wiley

Lenz, H.-J. (1998). *Multi-Data Sources and Data Fusion*. In: *Proc. New Techniques and Technologies for Statistics (NTTS) 1998*, EUROSTAT, pp. 139–146

Little R. J. A, Rubin D. B. (1987). „Statistical Analysis with Missing Data", Wiley & Sons, New York

Malvestuto, F.M. (1991). *Data Integration in Statistical Databases*. In: Michalewicz (ed.), *Statistical and Scientific Databases*, Chichester: Ellis Horwood, pp. 201–232

Oropallo, F., Orsini, M., Runci, M.C., „Software di supporto per la gestione delle liste delle unita' locali, per il monitoraggio delle attività di rilevazione e per la predisposizione dei primi risultati - Software to support census of firms: to manage local unit files, to monitor data collection, and to produce first outcomes". ISTAT, September 2001

Oropallo F. (2002). *Conceptual, logical and physical model of datamarts deliverable 2.1* - http://www.istat.it/diecofis/deliverable_list.htm.

Oropallo F., Skalbania, D. (2003). *Concept of IT framework issues and development of software for the creation of a multi-source data base - Analysis of the Software – deliverable 2.2* - http://www.istat.it/diecofis/deliverable_list.htm.

Oropallo F., Caruso E., (2003). *The management of DIECOFIS database - deliverable 2.3* - http://www.istat.it/diecofis/deliverable_list.htm.

Paass G., „Statistical match: Evaluation of existing procedures and improvements by using additional information", G.H.Orcutt and H.Quinke (eds) *Microanalytic Simulation Models to Support Social and Financial Policy*, Amsterdam: ElsevierScience, 1986, pp. 401-422

Porter, E., Winkler, W.E. (1997). *Approximate String Comparison and its Effect on an Advanced Record Linkage System*, RR97-02, U.S. Bureau of the Census, http://www.census.gov/srd/www/byyear.html.

Roberti P., Oropallo F., Inglese F., Lo Cascio L., de Martinis G. (2003) - Towards a Systemic Analysis of Italian Industrial Texture Review "Industria" 4/2002 – Il Mulino – November 2002

Roberti P., Oropallo F. (forthcoming) - Composite Indicators for the Measurement of Economic Performance - Productivity, Competitiveness and the New Information Economy - Business, Systemic and Measurement Issues, NESIS FP5 - ISTAT – Rome - June 26, 2003

Renssen R. H., „Use of Statistical Matching Techniques in Calibration Estimation", *Survey Methodology*, vol. 24, n. 2, 1998, pp. 171-183

Rodgers, W.L., „An Evaluation of Statistical Matching". *Journal of Business and Economic Statistics*, vol. 2, 1984, 91-102

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley & Sons, New York

Rubin, D.B., Belin, T.R., „Recent Developments in Calibrating Error Rates for Computer Matching", In: *Proc. 7th Annual Research Conference*, Washington, DC: U.S. Bureau of the Census, 1991, pp. 657–668

Ruggles, N., Ruggles, R., „A Strategy for Merging and Matching Microdata Sets", *Annals of Economic and Social Measurement,* vol. 3 (2), 1974, pp. 353–372

Scheuren, F., Winkler, W.E., „Regression Analysis of Data Files that are Computer Matched", *Survey Methodology,* vol. 19, 1993, pp. 39–58

Scheuren, F., Winkler, W.E., „Regression Analysis of Data Files that are Computer Matched II", *Survey Methodology* vol. 23, 1997, pp. 157–165

Schürmann, J. (1996). *Pattern Classification*, New York: John Wiley & Sons

Scotney, B.W., McClean, S.I., Rodgers, M.C., „Optimal and Efficient Integration of Heterogeneous Summary Tables in a Distributed Database", *Data and Knowledge Engineering,* vol. 29 (3), 1999, pp. 337–350

Schafer J.L., Olsen M. .K., „Multiple imputation for multivariate missing-data problems: a data analyst's perspective", *Multivariate Behavioral Research*, vol. 33, 1998, pp. 545-571

Sheth, A.P., Larson, J.A., „Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases", *ACM Computing Surveys,* vol. 22(3), 1990, pp. 183–236

Singh A. C., Mohl C. A., „Understanding Calibration Estimators in Survey Sampling", *Survey Methodology*, vol. 22, n.2, 1996, pp. 107-115

Singh, A.C, Mantel, H.J., Kinack, M.D., Rowe, G., „Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption", *Survey Methodology*, vol. 19, No.1, June 1993, pp. 59-79 - Statistics Canada

Sims, C.A., Comments on „Constructing a New Data Base From Existing Microdata Sets: The 1966 Merge File" by B.A. Okner, *Annals of Economic and Social Measurement* vol. 1, 1972, pp. 343–345

Winkler W. E. (1995). „Matching and Record Linkage", in B. G. Cox et al. (ed.), *Business Survey Methods*, Wiley & Sons, New York, pp. 920-935 (355-384)

Winkler, W.E., "Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage", In: *ASA, Proc. Section on Survey Research Methods*, 1993, pp. 274–279. Also available as RR93-12, Washington, DC: U.S. Bureau of the Census, Statistical Research Division, http://www.census.gov/srd/www/byyear.html.

Winkler, W.E., „Advanced Methods for Record Linkage". In: *ASA, Proc. Section on Survey Research Methods*, 1994, pp. 467–472

Winkler, W.E. (1999c). *The State of Record Linkage and Current Research Problems*. RR99-04, U.S. Bureau of the Census, Statistical Research Division, http://www.census.gov/srd/www/byyear.html.

Winkler, W.E., Thibaudeau, Y. (1991). *An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census. RR91-09*, U.S. Bureau of the Census, Statistical Research Division, http://www.census.gov/srd/www/byyear.html.

Wu, C.F.J. „On the Convergence Properties of the EM-Algorithm", *Annals of Statistics,* vol. 11 (1), 1983, pp. 95–103